FakeAVCeleb: A Novel Audio-Video Multimodal Deepfake Dataset

Supplementary Materials

Hasam Khalid[1], **Shahroz Tariq**[1], **Minha Kim**[1], **Simon S. Woo**[*,1,2,3]

[1] College of Computing and Informatics
[2] Department of Applied Data Science
[3] Department of Artificial Intelligence
Sungkyunkwan University, South Korea
{hasam.khalid, shahroz, kimminha, swoo}@g.skku.edu

A Dataset Publication

A.1 Links

We provide the following links to access our dataset: https://sites.google.com/view/fakeavcelebdash-lab/.
We provide the dataset through Google drive after receiving the Data Use Agreement (DUA) by Google form. DOI: http://doi.org/10.23056/FAKEAVCELEB_DASHLAB.

A.2 Hosting Platform

We host our dataset on Google Drive account, which belongs to DASH Lab managed by Simon S. Woo (corresponding authors of this paper) at Sungkyunkwan University, South Korea.

A.3 Access to Dataset

We have made a dataset request form to monitor and restrict the free use of our deepfake dataset (see Figure 10). As it has been suggested by experts that deepfake dataset can used by malicious actors to evade deepfake detectors. We have uploaded a small sample of our dataset on our GitHub. *Note: Everyone has to fill the dataset request form, which we will manually screen to limit misuse.*

A.4 Licensing

Our FakeAVCeleb dataset is available under Creative Commons 4.0 license and code is under MIT license (https://creativecommons.org/licenses/by/4.0/).

B Dataset Generation Methods

We used a total of 4 deepfake generation/synthesis methods. We will briefly explain each synthesis method below:

*Corresponding Author

35th Conference on Neural Information Processing Systems (NeurIPS 2021) Track on Datasets and Benchmarks.

Table 5: Summary of deepfake detection methods compared in this paper.

Dataset	Repositories	Release Data
UADFV [1]	https://bitbucket.org/ericyang3721/headpose_forensic	2018.11
DeepfakeTIMIT [2]	https://www.idiap.ch/en/dataset/deepfaketimit	2018.12
FF++ [3]	https://github.com/ondyari/FaceForensics	2019.01
Celeb-DF [4]	https://github.com/yuezunli/celeb-deepfakeforensics	2019.11
Google DFD [3]	https://github.com/ondyari/FaceForensics	2019.09
DeeperForensics [5]	https://github.com/EndlessSora/DeeperForensics-1.0	2020.05
DFDC [6]	https://ai.facebook.com/datasets/dfdc	2020.06
KoDF [7]	https://aihub.or.kr/aidata/8005	2021.06
FakeAVCeleb (Ours)	https://github.com/DASH-Lab/FakeAVCeleb	2021.12

FaceSwap [8] FaceSwap is a general deepfake generation method to swap faces between images or videos, retaining the body and environment context. We used Faceswap [8] software, an open-source face-swapping tool used to generate high-quality deepfake videos. Due to its popularity, it was used in FaceForensics++ datasets to generate the face-swapped dataset. The core architecture of this method consists of the encoder-decoder paradigm. A single encoder and two decoders (one for each source and target video) are trained simultaneously to build the face-swap model. The encoder extracts features from both videos while decoders reconstruct the source and target videos, respectively. The model is fed with frame-by-frame images of source and target video and trained for at least 80,000 iterations. We use this method because of its popularity as being a widely used deepfake generation method.

FSGAN [9] FSGAN is proposed by Nirkin et al. [9], which is the latest face-swapping method that has become popular recently. The key feature of this method is that it performs reenactment along with the face-swap. First, it applies reenactment on the target video based on the source video's pose, angle, and expression by selecting multiple frames from the source having the most correspondence to the target video. Then, it transfers the missing parts and blends them with the target video. This process makes it much easier to train and does not take much time to generate face-swapped video. We use the code from the official FSGAN GitHub repository [10]. We used the best quality swapping model recommended by the authors of FSGAN to prepare our dataset, by fine-tuning the input video pairs and generating better quality results. We adopt this method because of its efficiency and better quality of the results.

Wav2Lip [11] Recently, audio-based facial reenactment techniques along with lip-syncing have been proposed by researchers [11, 12]. In lip-sync, the source person controls the mouth movement, and in face reenactment, facial features are manipulated in the target video. One of the most recent audio-driven facial reenactment methods is Wav2Lip [11], which aims to lip-sync the video with respect to any desired speech signal by reenacting the face. Unlike LipGAN [12], which further fine-tuned the model on the generated frames, using a pretrained lip-sync discriminator to learn the lip-sync with respect to the desired audio accurately, Wav2Lip used five video frames and the respective audio spectrogram to capture the video's temporal context. We used this facial reenactment method because of the efficiency of its synthesis process for generating lip-synced video.

SV2TTS [13] Transfer Learning from Speaker Verification to Multispeaker Text-To-Speech Synthesis (SV2TTS) [13] is a real-time voice cloning (RTVC) tool that allows us to clone a voice from a few seconds of input audio. SV2TTS consists of three sub-models which are trained independently. First, an encoder network is trained on a speaker verification task. It generates a fixed-dimensional embedding vector of input audio, a synthesis network based on Tacotron 2 that generates Mel spectrogram, and a WaveNet-based vocoder network that converts the Mel spectrogram into time-domain waveform samples. SV2TTS works in real-time, taking the text and reference audio as input, and generating a cloned audio based on the input audio. We used this tool to generate cloned audios of our real video dataset.

B.1 Deepfake Detection Baseline Methods

We use eleven different deepfake detection methods to evaluate our FakeAVCeleb. We use these methods based on their code availability and frame-level AUC scores, where these methods analyze

individual frames and output a classification score. We use the default parameters provided with each compared detection method. We briefly discuss each of the detection methods below:

Capsule [14] This method is based in capsule structures to perform deepfake classification. This original model is trained on the FaceForensics++ dataset.

HeadPose [1] This method detects deepFake videos based on the inconsistencies in the head poses caused by deepfake generation methods. The model is based on SVM, and is trained on the UADFV dataset.

Visual Artifacts (VA) [15] This method detects videos based on visual artifacts in the eyes, teeth and facial areas in synthesized videos. Two variants of this model is provided: VA-LogReg and VA-MLP. And we used both of them to evaluate our dataset. VA-LogReg uses a simpler logistic regression model, while VA-MLP uses a feed forward neural network classifier. Both models are trained on real videos from CelebA dataset and fake videos from YouTube.

Xception [3] This baseline method is based on the XceptionNet model [16]. We use Xception in two different settings, Xception-raw (detection on raw frames) and Xception-comp15 (detection on compressed frames).

Meso4 [17] This method is based on Deep Neural Network (DNN) which targets mesoscopic properties of real and fake images. The model is trained on anonymous deepfake dataset. We use two variant of MesoNet, which are, Meso4 and MesoInception4.

F3Net [18] This deepfake detection method achieves the state-of-the-art performance by employing frequency-aware clues and local frequency statistic for deepfake detection in frequency domain.

Face X-ray [19] Li et al. [19] proposed a method that detect deepfakes by combining classification and segmentation, based on blending boundaries of manipulated images.

LipForensics [20] LipForensics is a recent work to detect facial forgeries based on unnatural lip movement by paying attention to the mouth area. They employ spatio-temporal network for the detection.

Multimodal-1 [21] Multimodal-1 was developed for the classification of food recipes. They use two neural networks to extract features from visual and textual modality, and them performs classification using a third neural network. To perform the classification on our FakeAVCeleb dataset, we modified the model with respect to the video and audio modalities. We removed the neural network for textual modality and replicated the neural network of visual modality to use this model on the multimodal dataset.

Multimodal-2 [22] Multimodal-2 is an open-source method that is developed for movie genre prediction. It takes movie poster (image) as input for a CNN block, movie genre (text) as input for an LSTM block, and then concatenate the output to perform classification. To use this model on our multimodal dataset, we removed the LSTM block and replaced it with the same CNN block, resulting in two CNN blocks, one for visual and one for audio modality.

CDCN [23] Central Difference Convolutional Networks (CDCN) [23] was developed to solve the task of face anti-spoofing. The model takes three-level fused features (low-level, mid-level, high-level) extracted for predicting facial depth. To perform the experiment, we modified the model by removing the third modality since it contains all three visual modalities.

Preprocessing We first preprocess the dataset before passing it to the models for training. As mentioned in main text, preprocessing was performed separately for videos and audios. Since we collected videos from VoxCeleb2 dataset, these videos are already face-centered and cropped. We extract respective frames from each video and store them separately, and then extract audios from the videos and store them in *.wav* format with a sampling rate of 16 kHz. Before inputting audio directly to the model for training, we first compute Mel-Frequency Cepstral Coefficients (MFCC) features by applying a $25ms$ Hann window [24] with $10ms$ window shifts, followed by a fast Fourier transform (FFT) with 512 points. As a result, we obtain a 2D array of 80 MFCC features ($D = 80$) per audio frame and store the resulting MFCC features as a three channel image, which is then passed to the model as an input to extract speech features so that it learns the difference between real and fake human voices.

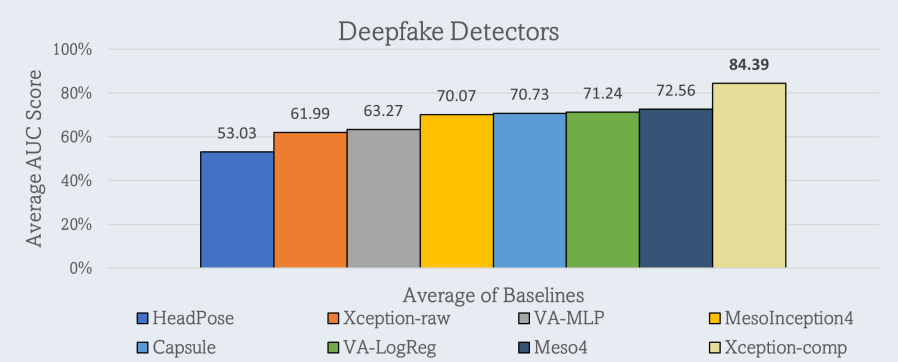


Figure 5: Average AUC score of deepfake detectors over all datasets.

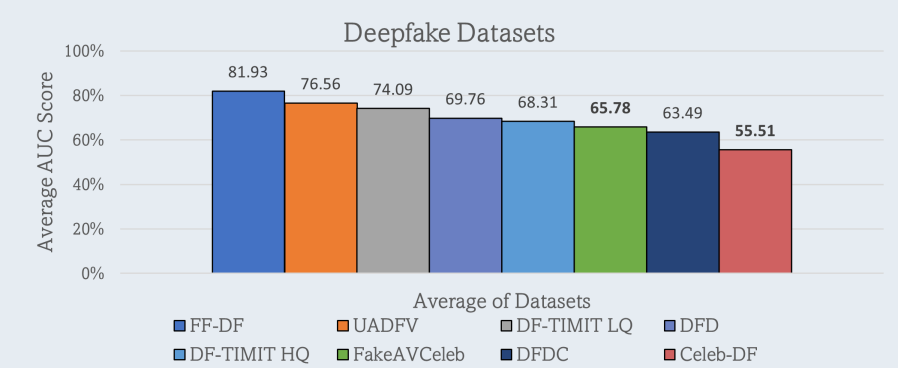


Figure 6: Average AUC score for each deepfake dataset on SOTA baseline methods.

B.2 Summary of Results

In Figure 5, we present the average of the AUC score for each baseline deepfake detector on eight different datasets, which are FF-DF, UADFV, DFD, DF-TIMIT LQ, DF-TIMIT HQ, FakeAVCeleb, DFDC, and Celeb-DF. We find that on average Xception-comp demonstrates the best (72.5%) and Headpose shows the worst (49.0%) detection performance.

In Figure 6, we present the average AUC score for each dataset based on eight different detection methods, which are Headpose, Xception-raw, Xception-comp, VA-MLP, VA-LogReg, MesoInception4, Meso4 and Capsule. The FF-DF dataset shows the highest detection score making it the easiest one to detect, while CelebDF shows the lowest, making it the hardest one to detect. The detection score for our FakeAVCeleb dataset is relatively close to CelebDF. However, our dataset has an additional advantage of having fake audio data.

C Additional Experiments

We also performed additional experiments on our dataset in unimodal, ensemble, and multimodal settings. The following sections cover the details of these experiments. *Note: The results of this work are based on the FakeAVCeleb v1.2 database. In the future, we will release new versions of FakeAVCeleb as the dataset's quality improves. Please visit our GitHub page*[2] *for the most recent results for each baseline on newer versions of FakeAVCeleb.*

[2] `https://github.com/DASH-Lab/FakeAVCeleb`

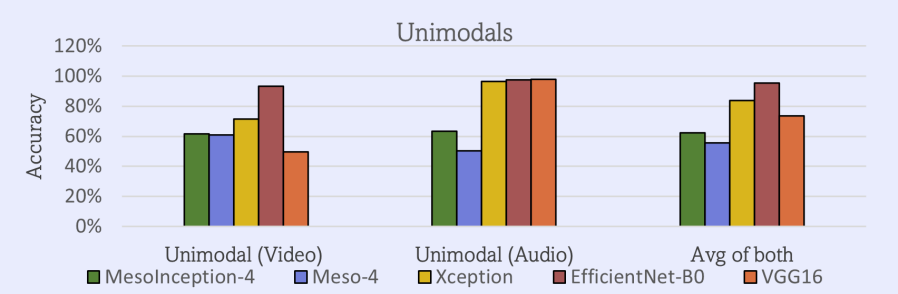


Figure 7: The AUC scores (%) of the five unimodal methods. We use three types of detection methods to evaluate the FakeAVCeleb dataset. We use a single modality, i.e., either $\mathcal{A}_{only}$ or $\mathcal{V}_{only}$, to train the model.

C.1 Unimodal Results

The performance of the baseline trained only on audio ($\mathcal{A}_{only}$) or only on video ($\mathcal{V}_{only}$) on the test set, which contains real and all three categories of fake videos from the FakeAVCeleb dataset, is described in this section.

The results of deepfake detection using unimodal baselines for $\mathcal{A}_{only}$ and $\mathcal{V}_{only}$ are presented in Figure 7. We can observe that the best AUC scores for $\mathcal{A}_{only}$ and $\mathcal{V}_{only}$ are approximately 97% and 93%, respectively.

C.2 Results for $\mathcal{V}_{only}$ Trained Classifier

In terms of video, as shown in Figure 7, EfficientNet-B0 [25] and VGG [26] achieves the performance of 93.3% and 49.6%, which are the best and lowest results of AUC score, respectively. In particular, Meso4's recall score suggests that the model fails to detect most deepfake videos ($\mathcal{V}_{\mathcal{F}}$). On the other hand, EfficientNet-B0 outperformed Xception [16] on this task, though Xception is the best performer on other deepfake datasets such as FaceForensics++ [3].

C.2.1 Results for $\mathcal{A}_{only}$ Trained Classifier

For audio, as shown in Figure 7, VGG achieves the best AUC score of 97.8%, and Meso4 [17] shows the lowest AUC score of 73.5%, which means that Meso4 overfit real class and fake class for audio detection, respectively. Moreover, we can observe that there are no baselines to provide satisfactory detection performance, indicating that SOTA deepfake detection methods are not suitable for deepfake audio ($\mathcal{A}_{\mathcal{F}}$) detection. The models developed for human speech verification or detection may perform better in detecting $\mathcal{A}_{\mathcal{F}}$. However, we have not considered such methods in this work. And, we expect that such methods can be explored for future work.

C.2.2 Summary of Unimodal Results

EfficientNet-B0 exhibits the most stable average performance of 95% for $\mathcal{A}_{only}$ and $\mathcal{V}_{only}$. Overall, the poor performance of SOTA deepfake detection models in Figure 7 indicates that the fake audios and videos in our dataset are of realistic quality, making it difficult for detectors to distinguish them from real ones.

C.3 Ensemble Results

In Figure 8, we used the SOTA unimodal baselines to make ensemble of unimodal $\mathcal{A}_{only}$ and unimodal $\mathcal{V}_{only}$ classifier. The ensemble network of EfficientNet-B0 performs the best (82.8%) as compared to an ensemble of Xception [3] classifier (51.4%) and F3Net [18] (47.6%) on the test set, respectively. Meanwhile, Meso4 shows the second best performance of 58.2% and MesoInception4 [19] ensemble achieves the third highest AUC score of 55.9%. On the other hand, Face X-ray [19] ensemble achieves mediocre AUC scores of 53.5%. Overall, as shown in Figure 8, we can observe that the choice of soft- or hard-voting did not have a significant impact on the performance of the ensemble classifier,

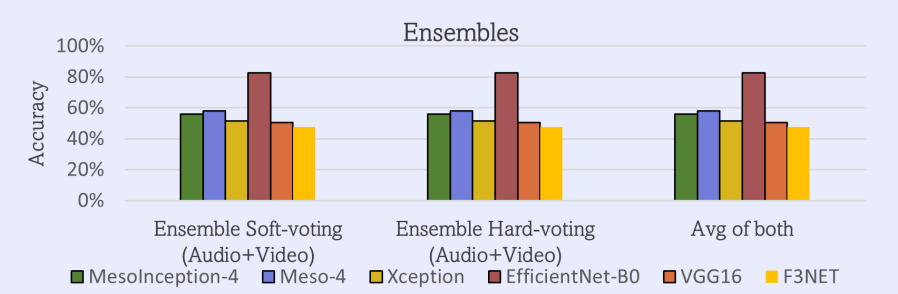


Figure 8: The AUC scores (%) of six models on our FakeAVCeleb which are used as an ensemble of two models (one for each modality), but trained separately. We use three types of evaluation settings, soft-voting, hard voting and average of both, to evaluate the dataset.

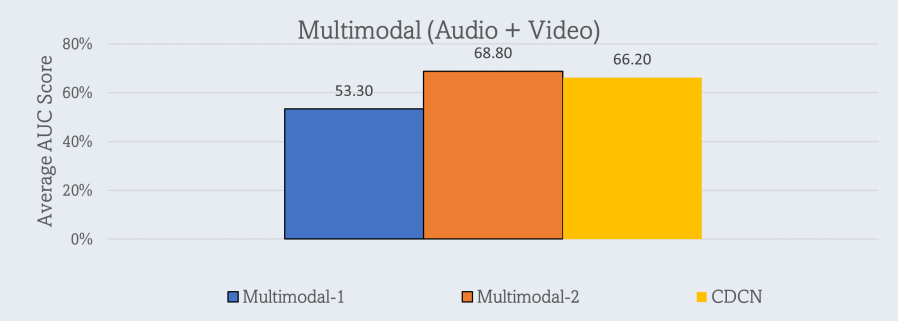


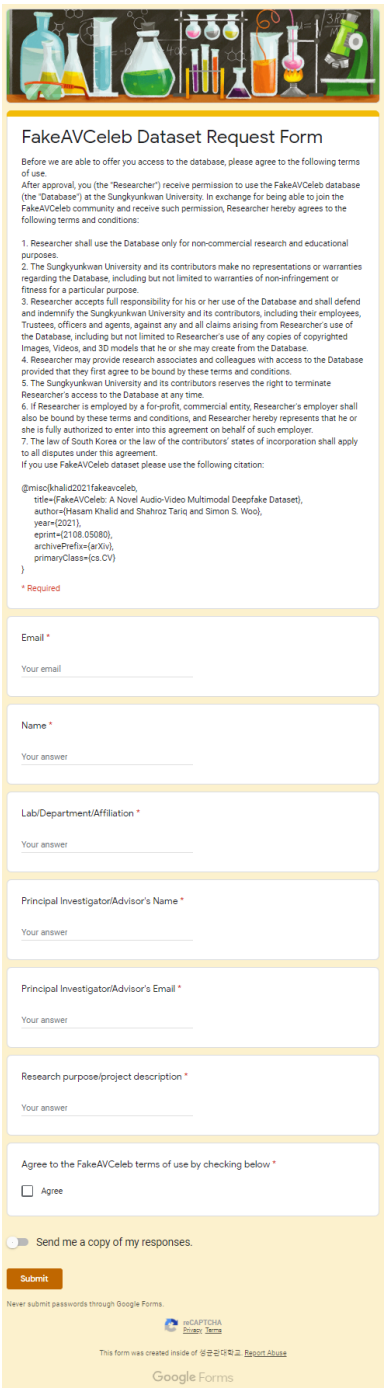
Figure 9: Multimodal detection performance ($\mathcal{V}$ and $\mathcal{A}$) on three different open source *multimodal* methods.

as they provide the similar prediction score. Note that this is because we have only two classifiers in our ensemble. Moreover, none of the ensemble-based methods could achieve a high detection score (i.e., > 90%), which represents that detecting multimodal $\mathcal{A}_{\mathcal{F}}\mathcal{V}_{\mathcal{F}}$ deepfakes is significantly harder, and more effective and advanced multimodal deepfake detection methods are required in the future.

C.4 Multimodal Results

We report on how the three baseline multimodals performed on the two modalities, $\mathcal{A}$ and $\mathcal{V}$, of the FakeAVCeleb multimodal dataset. The Multimodal-1 [21] was trained over 50 epochs. After selecting the best performing epoch, the model could classify with 53.3% AUC score. For Multimodal-2 [22], we trained on 50 epochs and evaluated them on our dataset, which shows 68.8% score. The third model, CDCN [23], was also trained on 50 epochs and provided 66.2% score. We can observe that Multimodal-1 and CDCN performed poorly compared to Multimodal-2. The possible reason for this result is that these models are designed to perform specific tasks, i.e., food recipe classification and movie genre prediction. Furthermore, the multimodal methods make it challenging to detect deepfakes when either the video is fake or the audio. Therefore, more research is needed in the development of multimodal deepfake detectors.

D Dataset Request Form



FakeAVCeleb Dataset Request Form

Before we are able to offer you access to the database, please agree to the following terms of use.
After approval, you (the "Researcher") receive permission to use the FakeAVCeleb database (the "Database") at the Sungkyunkwan University. In exchange for being able to join the FakeAVCeleb community and receive such permission, Researcher hereby agrees to the following terms and conditions:

1. Researcher shall use the Database only for non-commercial research and educational purposes.
2. The Sungkyunkwan University and its contributors make no representations or warranties regarding the Database, including but not limited to warranties of non-infringement or fitness for a particular purpose.
3. Researcher accepts full responsibility for his or her use of the Database and shall defend and indemnify the Sungkyunkwan University and its contributors, including their employees, Trustees, officers and agents, against any and all claims arising from Researcher's use of the Database, including but not limited to Researcher's use of any copies of copyrighted Images, Videos, and 3D models that he or she may create from the Database.
4. Researcher may provide research associates and colleagues with access to the Database provided that they first agree to be bound by these terms and conditions.
5. The Sungkyunkwan University and its contributors reserves the right to terminate Researcher's access to the Database at any time.
6. If Researcher is employed by a for-profit, commercial entity, Researcher's employer shall also be bound by these terms and conditions, and Researcher hereby represents that he or she is fully authorized to enter into this agreement on behalf of such employer.
7. The law of South Korea or the law of the contributors' states of incorporation shall apply to all disputes under this agreement.
If you use FakeAVCeleb dataset please use the following citation:

@misc{khalid2021fakeavceleb,
title={FakeAVCeleb: A Novel Audio-Video Multimodal Deepfake Dataset},
author={Hasam Khalid and Shahroz Tariq and Simon S. Woo},
year={2021},
eprint={2108.05080},
archivePrefix={arXiv},
primaryClass={cs.CV}
}

* Required

Email *

Your email

Name *

Your answer

Lab/Department/Affiliation *

Your answer

Principal Investigator/Advisor's Name *

Your answer

Principal Investigator/Advisor's Email *

Your answer

Research purpose/project description *

Your answer

Agree to the FakeAVCeleb terms of use by checking below *

Agree

Send me a copy of my responses.

Submit

Never submit passwords through Google Forms.

reCAPTCHA
Privacy Terms

This form was created inside of 성균관대학교. Report Abuse

Google Forms

Figure 10: A screenshot of Google form (`https://bit.ly/38prlVO`) to obtain access to FakeAVCeleb. If a researcher is using the FakeAVCeleb dataset, they can use the citation contained in Google form. The users must fill-in correct information, and should agree to follow the terms and conditions.

References

[1] Xin Yang, Yuezun Li, and Siwei Lyu. Exposing deep fakes using inconsistent head poses. In *ICASSP 2019-2019 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 8261–8265. IEEE, 2019.

[2] Conrad Sanderson and Brian C Lovell. Multi-region probabilistic histograms for robust and scalable identity inference. In *International conference on biometrics*, pages 199–208. Springer, 2009.

[3] Andreas Rossler, Davide Cozzolino, Luisa Verdoliva, Christian Riess, Justus Thies, and Matthias Nießner. Faceforensics++: Learning to detect manipulated facial images. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 1–11, 2019.

[4] Yuezun Li, Xin Yang, Pu Sun, Honggang Qi, and Siwei Lyu. Celeb-df: A large-scale challenging dataset for deepfake forensics. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 3207–3216, 2020.

[5] Liming Jiang, Ren Li, Wayne Wu, Chen Qian, and Chen Change Loy. Deeperforensics-1.0: A large-scale dataset for real-world face forgery detection. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 2889–2898, 2020.

[6] Brian Dolhansky, Joanna Bitton, Ben Pflaum, Jikuo Lu, Russ Howes, Menglin Wang, and Cristian Canton Ferrer. The deepfake detection challenge dataset. *arXiv preprint arXiv:2006.07397*, 2020.

[7] Patrick Kwon, Jaeseong You, Gyuhyeon Nam, Sungwoo Park, and Gyeongsu Chae. Kodf: A large-scale korean deepfake detection dataset. *arXiv preprint arXiv:2103.10094*, 2021.

[8] Iryna Korshunova, Wenzhe Shi, Joni Dambre, and Lucas Theis. Fast face-swap using convolutional neural networks. In *Proceedings of the IEEE International Conference on Computer Vision (ICCV)*, Oct 2017.

[9] Yuval Nirkin, Yosi Keller, and Tal Hassner. Fsgan: Subject agnostic face swapping and reenactment. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 7184–7193, 2019.

[10] Yuval Nirkin, Yosi Keller, and Tal Hassner. Fsgan - official pytorch implementation, 2020. [Online; accessed 31-May-2021]. URL: `https://github.com/YuvalNirkin/fsgan`.

[11] KR Prajwal, Rudrabha Mukhopadhyay, Vinay P Namboodiri, and CV Jawahar. A lip sync expert is all you need for speech to lip generation in the wild. In *Proceedings of the 28th ACM International Conference on Multimedia*, pages 484–492, 2020.

[12] Béatrice Daille, Éric Gaussier, and Jean-Marc Langé. Towards automatic extraction of monolingual and bilingual terminology. In *COLING 1994 Volume 1: The 15th International Conference on Computational Linguistics*, 1994.

[13] Ye Jia, Yu Zhang, Ron J. Weiss, Quan Wang, Jonathan Shen, Fei Ren, Zhifeng Chen, Patrick Nguyen, Ruoming Pang, Ignacio Lopez Moreno, and Yonghui Wu. Transfer learning from speaker verification to multispeaker text-to-speech synthesis, 2019. `arXiv:1806.04558`.

[14] Huy H Nguyen, Junichi Yamagishi, and Isao Echizen. Use of a capsule network to detect fake images and videos. *arXiv preprint arXiv:1910.12467*, 2019.

[15] Falko Matern, Christian Riess, and Marc Stamminger. Exploiting visual artifacts to expose deepfakes and face manipulations. In *2019 IEEE Winter Applications of Computer Vision Workshops (WACVW)*, pages 83–92. IEEE, 2019.

[16] François Chollet. Xception: Deep learning with depthwise separable convolutions. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 1251–1258, 2017.

[17] Darius Afchar, Vincent Nozick, Junichi Yamagishi, and Isao Echizen. Mesonet: a compact facial video forgery detection network. In *2018 IEEE International Workshop on Information Forensics and Security (WIFS)*, pages 1–7. IEEE, 2018.

[18] Yuyang Qian, Guojun Yin, Lu Sheng, Zixuan Chen, and Jing Shao. Thinking in frequency: Face forgery detection by mining frequency-aware clues. In *European Conference on Computer Vision*, pages 86–103. Springer, 2020.

[19] Lingzhi Li, Jianmin Bao, Ting Zhang, Hao Yang, Dong Chen, Fang Wen, and Baining Guo. Face x-ray for more general face forgery detection. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 5001–5010, 2020.

[20] Alexandros Haliassos, Konstantinos Vougioukas, Stavros Petridis, and Maja Pantic. Lips don't lie: A generalisable and robust approach to face forgery detection. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 5039–5049, 2021.

[21] xkaple01. Multimodal classification, 2019. [Online; accessed 31-May-2021]. URL: `https://github.com/xkaple01/multimodal-classification`.

[22] Dhruv Verma. Multimodal for movie genre prediction, 2021. [Online; accessed 31-July-2021]. URL: `https://github.com/dh1105/Multi-modal-movie-genre-prediction`.

[23] Zitong Yu, Chenxu Zhao, Zezheng Wang, Yunxiao Qin, Zhuo Su, Xiaobai Li, Feng Zhou, and Guoying Zhao. Searching central difference convolutional networks for face anti-spoofing. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 5295–5305, 2020.

[24] Wikipedia contributors. Hann function — Wikipedia, the free encyclopedia. `https://en.wikipedia.org/w/index.php?title=Hann_function&oldid=1001711522`, 2021. [Online; accessed 9-March-2021].

[25] Mingxing Tan and Quoc Le. Efficientnet: Rethinking model scaling for convolutional neural networks. In *International Conference on Machine Learning*, pages 6105–6114. PMLR, 2019.

[26] Karen Simonyan and Andrew Zisserman. Very deep convolutional networks for large-scale image recognition. *arXiv preprint arXiv:1409.1556*, 2014.